Building a Statistical Machine Translation System from Scratch: How Much Bang for the Buck Can We Expect?

Ulrich Germann
USC Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
germann@isi.edu

Abstract

We report on our experience with building a statistical MT system from scratch, including the creation of a small parallel Tamil-English corpus, and the results of a task-based pilot evaluation of statistical MT systems trained on sets of ca. 1300 and ca. 5000 parallel sentences of Tamil and English data. Our results show that even with apparently incomprehensible system output, humans without any knowledge of Tamil can achieve performance rates as high as 86% accuracy for topic identification, 93% recall for document retrieval, and 64% recall on question answering (plus an additional 14% partially correct answers).

1 Introduction

Crises and disasters frequently attract international attention to regions of the world that have previously been largely ignored by the international community. While it is possible to stock up on emergency relief supplies and, for the worst case, weapons, regardless of where exactly they are eventually going to be used, this cannot be done with multilingual information processing technology. This technology will often have to be developed after the fact in a quick response to the given situation. Multilingual data resources for statistical approaches, such as parallel corpora, may not always be available.

In the fall of 2000, we decided to put the current state of the art to the test with respect to the rapid construction of a machine translation system from scratch.

Within one month, we would

- hire translators;
- translate as much text as possible; and
- train a statistical MT system on the data thus created.

The language of choice was Tamil, which is spoken in Sri Lanka and in the southern part of India. Tamil is a head-last language with a very rich morphology and therefore quite different from English.

2 Data Collection and Preparation

2.1 Obtaining Tamil Data

Tamil data is not very difficult to find on the web. There are several Tamil newspapers and magazines with online editions, and the large international Tamil community fosters the use of the Internet for the dissemination of information. After initial investigation of several web sites we decided to download our experimental corpus from www.tamilnet.com, a news site that provides local news on Sri Lanka in both Tamil and English. The Tamil and English news texts on this site do not seem to be translations of each other. The availability of a fairly large in-domain corpus of local news on Sri Lanka in English (over 2 million words) allowed us to train an in-domain English language model of Sri Lankan news.

2.2 Encoding and Tokenization

Tamil is written in a phonematic, non-Latin script. Several encoding schemes exist in parallel. Even though the Unicode standard includes a set of glyphs for Tamil, it is not widely used in practice. Most web sites that offer Tamil language material assume Latin-1 encoding and rely on special true type fonts, which often are also offered for free download at those sites. Tamil text is therefore fairly easy to identify on web sites via the *face* attribute of the HTML *font* tag. All that is necessary is a list of Tamil font names used by the different sites, and knowledge about which encodings these fonts implement. While we could restrict ourselves to one data source and encoding for our experiment, any large-scale system would have to take this into account. In order to make the source text recognizable to humans who have no knowledge of Tamil, we decided to work with transliterated text[1].

2.3 Translating the Corpus

Originally we hoped to be able to create a parallel corpus of about 100,000 words on the Tamil side within one month, using several translators. Professional

[1]Translations, however, were produced from the original Tamil.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2001		**00-00-2001 to 00-00-2001**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Building a Statistical Machine Translation System from Scratch: How Much Bang for the Buck Can We Expect?	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Information Sciences Institute,University of Southern California,4676 Admirality Way Suite 1001,Marina del Rey,CA,90292**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**8**	

translation services in the US currently charge rates of about 30 cents per *English* word for translations from Tamil into English. Given that the English translation of a Tamil text usually contains about 1.2 times as many words as the Tamil original, the translation of a corpus of 100,000 Tamil words would cost approximately USD 36,000. This was far beyond our budget.

In India, by comparison, raw translations may cost as little as one cent per Tamil word[2]. However, outsourcing the translation work abroad was not feasible for us, since we had neither the administrative infrastructure nor the time to manage such an effort. Also, working with partners so remote would have made it very difficult to communicate our exact needs and to implement proper quality control.

We finally decided to hire as translators four entering and second-year graduate students in the department of engineering whose native language is Tamil and who had responded to an ad posted in the local mailing list for students from India.

In order to manage the corpus translation process, we set up a web interface through which the translators could retrieve source texts and upload their translations, post-editors could post-edit text online, the project progress could be monitored, and all incoming text was available to other project members as soon as it was submitted.

We originally assumed that translators would be able to translate about 500 words per hour if we were content with raw translations and hardly any formatting, and if we allowed them to skip difficult words or sentences. This estimate was based on an internal evaluation, in which multilingual members of our group translated sample documents from their native language (Arabic, German, Romanian) into English and kept track of the time they spent on this.

It turned out that our expectations were very much exaggerated with both respect to translation speed and the quality of translation. The actual translation speed for Tamil varied between 156 and 247 words per hour with an average of 170 words per hour. In 139 hours of reported translation time (over a period of eventually 6 weeks), about 24,000 words / 1,300 sentences of Tamil text were translated, at an effective cost of ca. 10.8 cents per Tamil word (translators' compensation plus administrative overhead). This figure does not include the effort for manually post-editing the translations by a native speaker of English (12-16 person hours).

The overall organization of the project (source data retrieval, hiring and management of the translators, design and implementation of the web interface for managing the project via the Internet, development of transliterator and stemmer, etc.) required an additional estimated 2.5 person months. However, a good part of this effort led to resources that can also be used for other purposes.

[2]Personal communications with Thomas Malten, University of Cologne.

2.4 Lessons Learned for Future Projects

If we were to give advice for future, similar projects, we would emphasize and recommend the following:

2.4.1 Good translators are not easy to find

It is difficult to find good translators for a short-term commitment. Unless one is willing to pay a premium price, it is unlikely that one will find professional translators who are willing to commit much of their time for a limited period of time and on short notice.

2.4.2 Make the translation job attractive

As foreign students, our translators would each have been allowed to work up to twenty hours per week. None of them did, because the work was frustrating and boring, and because they found more attractive, long term employment on campus. Our translators' frustration may have been fostered by several factors:

- the differences between Sri Lankan Tamil (the variety used in our corpus) and the Tamil spoken in Southern India (the native language of our translators), which made translating, according to our translators, very difficult;
- the lack of translation experience of our translators; and
- our high expectations. We originally told our translators that since they were not working on site, we would expect the translation of 500 words per hour reported. When we later switched to hourly pay regardless of translation volume, the translation volume picked up slightly.

2.4.3 Be prepared to post-edit

In professional translating, translators typically translate into their native language only. One may not be able to find translators with English as their native language for low density or "small" languages, so it may be necessary to have the translations post-edited by people with greater language proficiency in English.

2.4.4 Have translators and post-editors work on site

It is better to have translators and post-editors work on site and ideally as teams, so that they can resolve ambiguities and misunderstandings immediately without the delays of communicating indirectly, be it by email or other means. A post-editor who does not know the source language may misinterpret the translator, as the following case from our corpus illustrates:

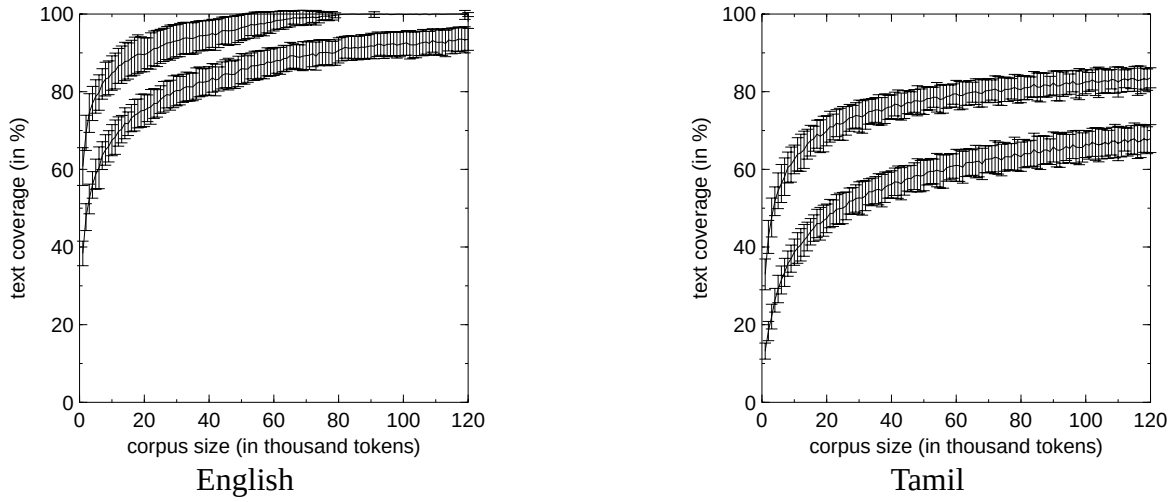

Figure 1: Text coverage on previously unseen text for English (left) and Tamil (right). The upper line in each graph shows the coverage by tokens that have been seen at least once, the lower line shows the coverage by tokens that have been seen at least 5 times. The error bars indicate standard deviation.

Raw translation: *Information about the schools in which people who migrated to Kudaanadu are staying is being gathered.*

Post-edited version: *Information about the schools in* (sic!) *which immigrants to Kudaanadu are attending is being gathered.*

In this case, the post-editor clearly misinterpreted the translator. What the translator meant to and actually did say is that information was being gathered about the schools in which migrants/war refugees who had arrived in Kudaanadu had found shelter. However, the post-editor interpreted the phrase *people who migrated to Kudaana* as describing *immigrants* and assumed that information was being gathered about their education rather than their housing.

3 Evaluation Experiments

3.1 A Priori Considerations

The richer the morphology of a language is, the greater is the total number of distinct word forms that a given corpus consists of, and the smaller is the probability that a certain word form actually occurs in any given text segment. Figure 1 shows the percentage of word forms in unseen text that have occurred in previously seen text as a function of the amount of previously seen text. The graph on the left shows the curves for English, the one on the right the curves for Sri Lankan Tamil. The graphs show the averages of 100 runs on different text fragments; the error bars indicate standard deviation.

The numbers were computed in the following manner: A corpus of 120,000 tokens was split into segments of 1000 tokens each. For each segment n_k, we computed how many of the tokens had been previously seen in the segments $n_1 \ldots n_{k-1}$. The upper line in the graphs shows the percentage of tokens in n_k that had occurred at least once before in the segments $n_1 \ldots n_{k-1}$, the lower line shows the percentage of tokens that had been seen at least five times before.

For the purpose of statistical NLP, it seems reasonable to assume that the lower curve gives a better indication of how many percent of previously unseen text we can expect to be "known" to a statistical model trained on a corpus of m tokens.

At a corpus size of 24,000 tokens, which is approximately the size of the parallel corpus we were able to create during our experiment, about 28% of all word forms in previously unseen Sri Lankan Tamil text cannot be found in the corpus, and 50% have been seen less than 5 times. In other words, if we train a system on this data, we can expect it to stumble over every other word! At a corpus size of 100,000 tokens, the numbers are 17% and 33%.

For English, the numbers are 9%/23% for a corpus of 24K tokens and 0%/8% for a corpus of 100K tokens.

In order to boost the text coverage we built a simple text stemmer for Tamil, based on the Tamil inflection tables in Steever (1990) and some additional inspection of our parallel corpus. The stemmer uses regular expression matching to cut off inflectional endings and introduce some extra tokens for negation and certain case markings (such as locative and genitive), which are all marked morphologically in Tamil. It should be noted that the stemmer is far from perfect and was only intended to be an interim solution. The performance increases are displayed in Figure 2. For a corpus size

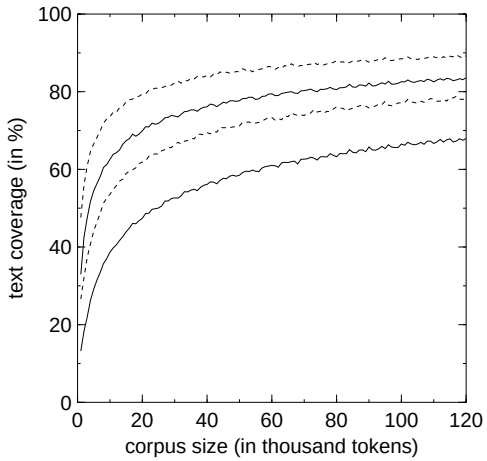

Figure 2: Text coverage increase by stemming for Tamil. The solid lines indicate text coverage for unstemmed data (seen at least once and at least five times, respectively), the dashed lines the text coverage for stemmed data.

of 24K tokens, the percentages of unknown items drop to 19% (from 28%; never seen before) and 36% (from 50%; seen less than 5 times). For a training corpus of 100K tokens, the numbers are 12% and 23% (from 17%/33%).

3.2 Task-Based Pilot Evaluation

Given these numbers, it is obvious that one cannot expect much performance from a system that relies on models trained on only 24K tokens of data. As a matter of fact, it is close to impossible to make any sense whatsoever of the output of such a system (cf. Fig. 3).

To get an estimate of the performance with more training data, we augmented our corpus with a parallel corpus of international news texts in Southern Indian Tamil which was made available to us by Fred Gey of the University of California at Berkeley (henceforth: *Berkeley corpus*). This corpus contains ca. 3,800 sentence pairs with 75,800 Tamil tokens after stemming (before stemming: 60,000; the difference is due to the introduction of additional markers during stemming). Some of the parallel data was withheld for system evaluation; the augmented training corpus (Berkeley and TamilNet corpus; short *B+TN*) had a size of 85K tokens on the Tamil side. The augmented training corpus had a text coverage of 81% (seen at least once; 75% without augmentation), and 67% (seen at least 5 times; 60% without augmentation), respectively, for Sri Lankan Tamil. We trained IBM Translation Model 4 (Brown et al., 1993) both on our corpus alone and on the augmented corpus, using the EGYPT toolkit (Knight et al., 1999; Al-Onaizan et al., 1999), and then translated a number of texts using different translation models and different transfer methods, namely glossing (replacing each Tamil word by the most likely candidate from the translation tables created with the EGYPT toolkit) and Model 4 decoding (Brown et al., 1995; Germann et al., 2001).

Figure 3 shows the output of the different systems in comparison with the human translation.

We then conducted the following experiments.

3.2.1 Document Classification Task

Seven human subjects without any knowledge of Tamil were given translations of a set of 15 texts (all from the Berkeley corpus) and asked to categorize them according to the following topic hierarchy:

- News about Sri Lanka
  - Reports about clashes between the Sri Lankan army and the Liberation Tigers
  - Sri Lankan security-related news (arrests, arms deals, etc.)
  - Sri Lankan political news (strikes, transport, telecom)
  - concerns Sri Lanka but doesn't fit any of the above
- News about Pakistan/India
  - Nuclear tests in Pakistan and India, including their aftermath (international reactions, etc.)
  - Corruption investigation against Benazir Buto
  - News about Pakistan/India but none of the above
- International news
  - Disasters, accidents
  - Nelson Mandela's birthday
  - Other international news
- Impossible to tell

Except for one duplicate set, each subject received a different set of translations. The sets differed in training parameters and the translation method used. Table 1 shows the results of this evaluation. The difference between the subjects 5a and 5b, who received the same set of translations, suggests that the individual classifiers' accuracy influences the results so much as to blur the effect of the other parameters. There seems to be a tendency for glossing to work better than Model 4 decoding. *Glossing*, in our system, is a simple base line algorithm that provides the most likely word translation for each word of input. Translation candidates and their probabilities are retrieved from the translation table, which is part of the translation model trained on the parallel corpus.

The document classification test is foremost and above all a measure of the quality of the translation table for frequently occurring words. In practice, actual

	Small Corpus 24K tokens (Tamil) of training data		Augmented Corpus 85K tokens (Tamil) of training data	
Human translation	**Gloss**	**Model 4 decoding**[a]	**Gloss**	**Model 4 decoding**
Government - United National Party meeting will not take place tomorrow.	vanni united national party kalloya tomorrow naTaipeRamaaTTa	children united national party tomorrow .	government united national party meet day naTaipeRamaaTTa	the government with united national party meets day .
The proposed meeting tomorrow, Thursday, between the Peoples Front Government and the United National Party regarding the New Political Document was announced as postponed to a later date.	friday progressive it and united national same kaTcikkumiTai tomorrow thursday naTaipeRaviru new political vaappu information kalloya piRitoru tin2attiRkupa pin2pooTappaTT-uLLataaka found is	on friday health mothers united national tomorrow on thursday new political found on information .	throughout progressive government the united national in kaTcikkumiTai day thursday leno new political vaappu found meet piRitoru tin2attiRkupa pin2pooTappaTT-uLLataaka movies .	throughout the government and the united national party . day thursday leno new political found meet movies .
Presidential Secretariat sources say that this meeting was postponed as the Sri Lankan President Chandrika Bandaranayakka has gone to a foreign country.	of lankan to and freedom tamilian now veLinaaTTukkuc cen2Riruppat iccantippu piRpooTappaTT-uLLataaka to secretariat in * is	’s lanka to advisor freedom of the tamilian team to secretariat is called minister .	of lankan president chandrika sankari freedom now veLinaaTTukkuc cen2Riruppat iccantippu piRpooTappaTT-uLLataaka president secretariat circles reported .	the sri lankan forces of president chandrika sankari freedom presently secretariat circles reported .
At the same time it is to be noted that the meeting of the sub committee examining the political document between the United National Party and the Government was held yesterday Tuesday evening	that about returning and iccantippu naTaipeRumen2a and ivvaTTaaram more and * is	her about the military and they were announced .	he countries returning the iccantippu naTaipeRumen2a the ivvaTTaaram more the reported .	his country and the returning to increase the radio .

[a] Brown et al. (1993); Brown et al. (1995)

Figure 3: Sample output of various systems

Table 1: Results of the Document Classification Task. Test subjects were asked to classify the translations of 15 documents into 4 major and 11 minor categories.

	input	pegging[a]?	transfer	correct	partially correct[b]	incorrect
1	raw	no	M4 decoding[c]	7	4	4
2	stemmed	yes	M4 decoding	8	3	4
3	stemmed	no	M4 decoding	13	2	0
4	raw	no	gloss	13	1	1
5a	stemmed	yes	gloss	8	3	4
5b	stemmed	yes	gloss	12	2	1
6	stemmed	no	gloss	11	2	2

[a] *pegging* causes the training algorithm to consider a larger search space
[b] correct top level category but incorrect sub-category
[c] translation by maximizing the IBM Model 4 probability of the source/translation pair (Brown et al., 1993; Brown et al., 1995)

classification might be performed by automatic procedures rather than humans. If we dare to accept the top performances of our human subjects as the tentative upper bound of what can be achieved with the current system using a translation model trained on 85K tokens of Tamil text and the corresponding English translations, we can conclude that the classification accuracy can exceed 86% (13/15) for fine-grained classification and reach 100% for coarse-grained classification. However, given the extremely small sample size in this evaluation, the evidence should not be considered conclusive.

3.2.2 Document Retrieval Task

The document retrieval task and the question answering task (see below) were combined into one task. The subjects received 14 texts (from the TamilNet corpus) and 15 lead questions plus 13 additional follow-up questions. Their task was to identify the document(s) that contain(s) the answer to the question and to answer the questions asked. Typical lead questions were questions such as *What is the security situation in Trinconmalee?*, or *Who is S. Thivakarasa?*; typical follow-up questions were *Who is in control of the situation?*, *What happened to him/her?*, or *How did (other) people react to what happened?*. As in the previous experiment, each subject received the output of a different system.

Table 2 shows the result of the document retrieval task. Again, the sample size was too small to draw any final conclusions, but our results seem to suggest the following. Firstly, the test subject in the group dealing with output of systems trained on the bigger training set tend to perform better than the ones dealing with the results of training on less data. This suggests that the jump from 24K to 85K tokens of training data might improve system performance in a significant manner. We were surprised that even with the poor translation performance of our system, recalls as high as 93% at a precision of 88% could be achieved. Secondly, the data shows that gaps are not randomly distributed over the data, but that some questions clearly seem to have been more difficult than others. One of the particular difficult aspects of the task was the spelling of names. Question 11, for example, asked *What happened to Chandra Kumar Abayasingh?*. In the translations, however, it was rendered in simple transliteration: *cantirakumaara apayacingka*. It requires a considerable degree of tenacity and imagination to find this connection.

3.2.3 Question Answering Task

In order to measure the performance in the question answering part of this evaluation, we considered only questions relevant to the documents that the test subjects had identified correctly. Because of the difficulty of the task, we were lenient to some degree in the evaluation. For example, if the correct answer was *the former president of the teacher's union* and the answer given was *an official of the teacher's union*, we still counted this as "close enough" and therefore correct. In addition, we also allowed partially correct answers, that is, answers that went into the right direction but were not quite correct. For example, if the correct answer was *The army imposed a curfew on fishing*, we counted the answer *the army is stopping fishing boats* as partially correct. All in all, it was very difficult to evaluate this section of the task, because it was often close to impossible to determine whether the answer was just an educated guess or actually based on the text. There were some cases where answers were partially or even fully correct even though the correct document had not been identified. In retrospect we conclude that it would have been better to have the test

Table 2: Recall and precision on the document retrieval task. Test subjects were asked to identify the document(s) containing the answers to 15 lead questions. Black dots indicate successful identification of at least one document containing the answer.

	training corpus	**transl. method**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**recall**	**precision**
1	TN[a]	glossing	•		•	•	•			•	•	•		•	•	•		67%	79%
2	TN	M4dec.[b]	•		•	•	•			•	•	•			•			53%	80%
3	TN	both[c]	•		•	•	•		•	•	•	•			•			60%	48%
4	TN[d]	both	•	•		•			•	•	•	•		•		•	•	67%	79%
5	B+TN.1	glossing	•	•	•	•	•		•	•	•	•		•	•		•	80%	87%
6	B+TN.1	M4dec.	•		•	•	•			•	•	•		•	•		•	67%	86%
7	B+TN.1	both	•		•	•	•			•	•	•	•	•	•	•	•	80%	86%
8	B+TN.1[f]	both	•	•			•			•	•			•	•		•	60%	73%
9	B+TN.1[g]	both	•	•	•	•	•		•	•	•	•	•	•	•	•	•	93%	88%
10	B+TN.2[h]	both	•	•	•		•	•	•	•	•	•		•	•	•	•	87%	75%
	Human Translations		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	100%	100%

[a]TamilNet corpus only; stemmed; 1291 aligned text chunks; 23,359 tokens on Tamil side; 1000 training iterations.
[b]IBM Model 4 decoding.
[c]Both glossing and IBM Model 4 decoding were available to the test subject.
[d]same as above, but trained with pegging option (more thorough search during training); 10 training iterations.
[e]Berkeley and TamilNet corpora; 5069 aligned text chunks; 85421 tokens on Tamil side; 100 training iterations.
[f]same as above; 10 training iterations.
[g]same as above, trained with pegging option; 10 training iterations.
[h]Berkeley and TamilNet Corpora, raw (unstemmed); 64439 tokens on Tamil side, 50 training iterations.

subjects mark up those text passages in the text that justify their answers.

Again, the data suggests that the difference in training corpus size does affect the amount of information that is available from the system output. Subjects using output of a system based on a translation model that was trained on only the TamilNet data tend to perform worse than subjects using output from a system based on a translation model trained on the larger corpus. The poor performance on test set No. 6 may suggest that for this task and at this level of translation quality, glossing provides more informative output than Model 4 decoding. This result is not particularly surprising, since we noticed that Model 4 decoding tends to leave out more words than acceptable. Clearly, this is one area where the translation model has to be improved.

Test set 10 is the only set produced by a system using a translation model trained on raw, unstemmed data. It is unclear whether the poor performance on question answering for this test set is due to a principally worse translation quality or the (lack of) tenacity and willingness of the test subject to work her way through the system output.

All in all, we were astonished by the amount of information that our test subjects were able to retrieve from the material they received (the top recall for the question answering task is 64%, plus an additional 14% partially correct answers!). However, using a system such as the one discussed in the paper is not an option for actual information processing. Especially those subjects that had to deal with the output of systems trained on the smaller corpus experienced the task as utterly frustrating and would not want to do it again.

Table 3: Accuracy on question answering. The test sets are the same as in Table 2. Only questions concerning documents that were identified correctly were considered in this evaluation.

test set	**training corpus**	No. of relevant questions	correct	partially correct
1	TN	17	2 = 12%	0 = 0%
2	TN	16	3 = 19%	4 = 25%
3	TN	18	2 = 11%	0 = 0%
4	TN	17	7 = 41%	5 = 19%
4	B+TN.1	23	14 = 61%	6 = 26%
6	B+TN.1	19	5 = 26%	3 = 16%
7	B+TN.1	22	14 = 64%	3 = 14%
8	B+TN.1	19	12 = 63%	1 = 5%
9	B+TN.1	26	14 = 54%	5 = 19%
10	B+TN.2	24	8 = 33%	9 = 38%
	human	28	24 = 86%	2 = 7%

4 Conclusions

We have reported on our experience with rapidly building a statistical MT system from scratch. Within ca. 140 translator hours, we were able to create a parallel corpus of about 1300 sentence pairs with 24,000 tokens on the Tamil side, at an average translation rate of approximately 170 Tamil words per hour.

Very clearly, the effort needed to create parallel data is one of the biggest obstacles to the rapid development of statistical MT systems for new languages.

With the output of a system which uses a translation model trained on the small amount of parallel data that we created during the course of our experiment, human test subjects achieved a recall of over 50% on the document retrieval task but generally performed poorly on question answering (less than 20%).

The addition of an additional corpus of 3,800 sentence pairs allowed us to estimate the benefits of increasing the overall corpus size by roughly 300%. Based on our experience with translating the TamilNet corpus, this additional effort would require an additional 450 translator and 36 to 48 post-editor hours.

With the additional training data, we were able to produce output that increased the performance on our evaluation tasks (document retrieval and question answering) to up to 93% for document retrieval and 64% for question answering.

With respect to the scenario of "MT in a month", we can now make the following calculation: If we assume that the average translator translates at a rate of 170 words/hour and is able to spend 6-7 hours per day on actual translations, then a translator can translate about 1000-1200 words per day. In order to translate a corpus of 100,000 words within one month (assuming a five-day work week), we therefore need **four to five full time translators**. For this effort, we can expect a translation system whose performance resembles the one shown in our evaluation.

This, of course, raises the following questions, which we are only able to ask but not to answer at this point.

- Can the translation model and the algorithms for statistical training be improved so that they require less data to produce acceptable results?
- Are there more efficient uses of scarce resources (such as language experts and translators) for building a statistical (or any other) MT system quickly, for example the creation of less but more informative data, e.g. a parallel corpus with alignments on the word level, or the compilation of a glossary/dictionary of the most frequently used terms?
- How do the various approaches compare with respect to the ratio of construction effort versus performance improvement when the MT systems are scaled up? One approach may show rapid improvements initially but also reach a plateau quickly, whereas another may show slow but steady improvements.
- Is there any potential for bootstrapping the resource creation process by using knowledge that can be extracted from little and poor data to speed up the creation of more and better data?

These are some of the the questions that will need to be addressed in future research on *Quick MT*.

5 Acknowledgments

This research has been funded by the DARPA TIDES program under grant No. N66001-00-1-8914. We would like to thank our translators as well as Fred Gey of the University of California at Berkeley and Thomas Malten of the University of Cologne for their kind support of our investigations into Tamil, and our test subjects for their tenacity and patience during this experiment.

References

Yaser Al-Onaizan, David Purdy, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Noah A. Smith, Franz Josef Och, and David Yarowsky. 1999. Statistical machine translation. Final report, Center for Language and Speech Processing, John Hopkins University.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Peter Brown, John Cocke, Stephen Della Pietra, Vincent Della Pietra, Frederick Jelinek, Jennifer Lai, and Robert Mercer. 1995. Method and system for natural language translation. U.S. Patent 5,477,451, Dec 19.

Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada. 2001. Fast decoding and optimal decoding for MT. In *39th Annual Meeting of the Association for Computational Linguistics (ACL-2001)*.

Kevin Knight, Yaser Al-Onaizan, David Purdy, Jan Curin, Michael Jahr, John Lafferty, Dan Melamed, Noah Smith, Franz Josef Och, and David Yarowsky. 1999. EGYPT: a statistical machine translation toolkit. `http://www.clsp.jhu.edu/ws99/projects/mt/`.

Sanford B. Steever. 1990. Tamil and the Dravidian languages. In Bernard Comrie, editor, *The World's Major Languages*, pages 725–746. Oxford University Press, New York, Oxford.